Verb Polarity Frames: a New Resource and its Application in Target-specific Polarity Classification

Klenner, Manfred ; Amsler, Michael ; Hollenstein, Nora

Verb Polarity Frames: a New Resource and its Application in Target-specific Polarity Classification

Manfred Klenner
Computational Linguistics
University of Zurich
Switzerland
klenner@cl.uzh.ch

Michael Amsler
Computational Linguistics
University of Zurich
Switzerland
mamsler@cl.uzh.ch

Nora Hollenstein
Computational Linguistics
University of Zurich
Switzerland
hollenstein@cl.uzh.ch

Abstract

We discuss target-specific polarity classification for German news texts. Novel, verb-specific features are used in a Simple Logistic Regression model. The polar perspective a verb casts on its grammatical roles is exploited. Also, an additional, largely neglected polarity class is examined: controversial texts. We found that the straightforward definition of 'controversial' is problematic. More or less balanced polarities in a text are a poor indicator of controversy. Instead, non-polar wording helps more than polarity aggregation. However, our novel features proved useful for the remaining polarity classes.

1 Introduction

We focus on fine-grained sentiment analysis in a document-level, target-specific polarity classification task. By fine-grained we refer to a sentiment analysis that captures sentiment composition at the phrase or even clause level based on reliable lexical resources, e.g., polarity lexicons. The task includes the recognition of targets and whether a (nearby) polar expression relates to it and how. Existing approaches have focused on different aspects of this task: the identification of targets and their components (Popescu and Etzioni, 2005), the induction of contextual polarity (Wilson et al., 2005), subjectivity word sense disambiguation (Akkaya et al., 2009), sentence-level composition (Moilanen and Pulman, 2007), and the specification of fine-grained lexical resources that help to better distinguish between factual and subjective language or even relate the polarity of expressions to emotion categories (Neviarouskaya et al., 2009). While recent research relying on a recursive neural tensor network (Socher et al., 2013) has shown that a high scoring sentiment analysis system that even copes with some effects and scopes of negation and with compositionality can also be trained with machine learning techniques, such an approach relies heavily on the annotated resources available, a sentiment treebank in this case.

Moving from English to other languages (German, in our case) confronts one with the lack of comparable resources, be it fine-grained polarity lexicons or – more seriously – the lack of gold standard data for training and evaluation of machine learning approaches. In order to change this situation, we have started to create a fine-grained polarity lexicon and a verb resource similar but not identical to (Neviarouskaya et al., 2009). We have also implemented a fast system carrying out sentiment composition, but one problem remained: how to evaluate in the absence of a (phrase- and sentence-level) gold standard[1]. Fortunately, we have access to a large text corpus (80,000 texts) where newspaper texts and dedicated actors in them are classified as positive,

[1]The MLSA corpus (Clematide et al., 2012) could have been a starting point, but is small and only captures NP-level composition.

negative, neutral or controversial[2]. This way, an extrinsic, i.e., application-oriented evaluation was possible. The goal was to reproduce the human-annotated target-specific classifications on the basis of our newly created resources. Could such a system be used to filter the huge amount of daily upcoming texts in order to, e.g., more directly access interesting (positive, negative, neutral or controversial) texts on a given target? The disadvantage of this resource is that it requires a demanding classification task, namely target classification including a class "controversial". There are only few approaches trying to cope with that problem (e.g. (Tsytsarau et al., 2010)). However, to withdraw these texts from our corpus was no option, since it would have made the intended application impossible. Unfortunately, no interannotator agreement (IAA) was measured for the text corpus. Thus, we conducted a small study (200 texts) in order to find out how well human annotators could reproduce the demanding "controversial" (expected) gold standard classifications. IAA turned out to be surprisingly low: if we take human performance as an upper bound, our system must beat 33% precision) – a poor value (overall accuracy was 66%).

In the present study, we combine text classification and features derived and aggregated from sentiment composition in an extrinsic evaluation in order to evaluate the impact of our newly created resources. No special attention was paid to "controversial target recognition". We not only believe that this task needs special treatment (as we argue in section 2), but also that no conclusions can be drawn given a gold standard class that even humans cannot reliably reproduce. (In addition to this, it should be mentioned that target specific sentiment analysis is considered to be more difficult in news texts than in other text genres (Balahur et al., 2010)).

The rest of this paper is structured as follows. After the related work section, we briefly discuss the origin and intended usage of our text corpus, introduce our resources and describe our approach to sentiment composition. In the experimental sections, we describe and measure the impact of our various features, given different partitions (and, thus, class distributions) of our text corpus. Finally, we draw some conclusions.

[2]No fine-grained annotations are available, e.g. no phrase- or sentence-level polarities.

2 Related Work on Controversial Texts

There are only a few approaches dealing with the classification of controversial targets. In (Choi et al., 2010) and (Tsytsarau et al., 2010), the hypothesis is that a topic is controversial if the difference between negative and positive phrase level polarity is within a heuristically determined range. This is in line with the annotation guidelines of our gold standard corpus, where target evaluations are considered controversial if positive and negative aspects are balanced and no polarity clearly prevails. We have included features capturing positive-negative ratios of various types of polar expressions (lexicon-based, composition-based etc.) in our experiments - without success.

(Choi et al., 2010) try to detect (new) controversial topics (and subtopics) from text collections, while we focus on intra-text detection of controversial discussions.

Dori-Hacohen and Allan (2013) try to find out if a web page discusses a (known) controversial topic. A web page is controversial if it is similar to a controversial Wikipedia article (on that topic).

3 Text Corpus

Our text corpus, used as a gold standard, was created by the *fög* institute (Research Institute for the Public Sphere and Society)[3] carrying out quantitative-qualitative media content and media reputation analysis. This institute analyses the media reputation of the key sectors of financial and real economies. Media reputation is defined by ((Deephouse, 2000), p. 1097) as "the overall evaluation of the firm presented in the media resulting from the stream of media stories about the firm". The content analysis examines how frequently and strongly (centrality) the media report on specific companies and how they were evaluated (polarity). The recorded encodings (positive, neutral, negative and controversial) allow the institute to build a Media Reputation Index.

[3]*http://www.foeg.uzh.ch/*

4 Fine-grained Polarity Lexicon

We aim at a compositional treatment of phrase- and sentence-level polarity. In order to assure high quality, we rely on a manually crafted polarity lexicon specifying the polarities of words (not word senses). Recently, fine-grained distinctions have been proposed that distinguish between various forms of positive and negative polarities, e.g. (Neviarouskaya et al., 2009). For instance, the appraisal theory (Martin and White, 2005) suggests to distinguish between appreciation ("sick friend"), judgement ("deceitful friend") and emotion ("angry friend") . Especially if polarity composition comes into play, it might be crucial to keep these different kinds of polarity separate. We want to properly distinguish cases like "admire a sick friend" (no polarity expectation conflict) from "admire a deceitful friend" - where a polarity conflict occurs (in general, "admire" expects a positive direct object, however a factually negative NP with a non-active connotation does not seem to violate this condition).

We have adopted the categories of the appraisal theory. Our German polarity lexicon comprises about 7,000 single-word entries (nouns, adjectives, adverbs), manually annotated for positive and negative prior polarity where each class further specifies whether a word is factually, morally or emotionally polar. We also coded whether the word involves an active part of the related actor (where applicable) and whether it is weakly or strongly polar. Our ultimate goal is to combine this resource with our verb resource (described in section 5.2) in order to predict the polarity of the arguments of a verb or even to be able to deal with conflicts arising from violated polarity expectations of the verb. In the present study, we use the fine-grained polar values from the lexicon as features (e.g. we count how many words with a prior polarity from the factual axis appear together with the target). But we also enumerate the number of positive and negative arguments stemming from verb expectations and effects (see next section).

Also part of our lexicon are shifters (inverting the polarity, e.g., "a good idea" (positive) vs. "no good idea" (negative)), intensifiers and diminishers.

5 Sentiment Composition

5.1 Phrasal Level

According to the principle of compositionality and along the line of other scholars (e.g. (Moilanen and Pulman, 2007)), after mapping polarity from the lexicon to the words of the text, in the next step we calculate the polarity of nominal and prepositional phrases, i.e., based on the lexical marking and taking into account syntactic (dependency) structure, we conduct a composition of polarity for the phrases.

In general, the polarities are propagated bottom-up to their respective heads of the NPs/PPs in composition with the other subordinates. To conduct this composition we convert the output of a dependency parser (Sennrich et al., 2009) into a constraint grammar format and use the `vislcg3`-tools (VISL-group, 2014) which allows us to write the compositional rules in a concise manner.

5.2 Verb Polarity Frames: Effects and Expectations

In order to merge the polar information of the NPs/PPs on the sentence level one must include their combination via their governor which is normally the verb. Neviarouskaya et al. (2009) propose a system in which special rules for verb classes relying on their semantics are applied to attitude analysis on the phrase/clause-level. Reschke and Anand (2011) show that it is possible to set the evaluativity functors for verb classes to derive the contextual evaluativity, given the polarity of the arguments. Other scholars carrying out sentiment analysis on texts that bear multiple opinions toward the same target also argue that a more complex lexicon model is needed and especially a set of rules for verbs that define how the arguments of the subcategorization frame are affected - in this special case concerning the attitudes between them (Maks and Vossen, 2012).

Next to the evidence from the mentioned literature and the respective promising results, there is also a strong clue coming from error analysis concerning sentiment calculation in which verbs are treated in the same manner as the composition for polar adjectives and nouns described above. This shows up especially if one aims at a target

specific (sentence-level) sentiment analysis: in a given sentence "*State attorney X accuses Bank Y of investor fraud.*" one can easily infer that *accuse* is a verb carrying a negative polarity. But in this example the direct object *Bank Y* is accused and should therefore receive a negative "effect" while the *State attorney X* – as the subject of the verb – is not negatively affected (it is his duty to investigate and prosecute financial fraud). Second, the PP *of investor fraud* is a modification of the accusation (giving a reason) and there is intuitively a tendency to expect a negative polarity of this PP - otherwise the accusation would be unjust (In the example given, the negative expectation matches with the composed polarity stemming from the lexically negative "fraud"). So it is clear that the grammatical function must be first determined in order to accurately calculate the effects and expectations that are connected to the lexical-semantic meaning of the verb.

Furthermore, the meaning of the verb (and therefore the polarity) can change according to the context (cf. "report a profit" (positive) vs. "report a loss" (negative) vs. "report an expected outcome"(neutral)). This leads to a conditional identification of the resulting verb polarity (or verbal phrase respectively) in such a manner that the polarity calculated for the head of the object triggers the polarity of the verb. In German, for instance, there are verbs that not only change their polarity in respect to syntactic frames (e.g. in reflexive form) but also in respect to the polarity of the connected arguments, too (see Tab. 1). Of course, any further modifiers or complements of the verb must also be taken into account.

German	English	Polarity
für die Kinder sorgen	to take care of the kids	positive
für Probleme[neg.] sorgen	to cause problems	negative
für Frieden[pos.] sorgen	to bring peace	positive
sich sorgen	to worry	negative

Table 1: Several examples for the use of the German verb "sorgen".

We therefore encode the impact of the verbs on polarity concerning three dimensions: effects, expectations and verb polarity. While effects should be understood as the outcome instantiated through the verb, expectations can be understood as anticipated polarities induced by the verb. The verb polarity as such is the evaluation of the whole verbal phrase. To sum up: in addition to verb polarity, we introduce effects and expectations to verb frames which are determined through the syntactic pattern found (including negation), the lexical meaning concerning polarity itself and/or the conditional polarity respective to the bottom-up calculated prevalent polarities. This results at the moment in over 120 classes of verb polarity frames with regard to combinations of syntactic patterns, given polarities in grammatical functions, resulting effects and expectations, and verb polarity.

As an example we take the verb class *fclass_subj_neg_obja_eff_verb_neg* which refers to the syntactic pattern (subject and direct object) and at the same time indicates which effects and/or expectations are triggered (here negative effect for the direct object). If the lemma of the verb is found and the syntactic pattern is matched in the linguistic analysis, then we apply the rule and assign the impacts to the related instances. However, the boundary of syntax is sometimes crossed in the sense that we also include lexical information if needed. For instance, if we specify the lemma of the concerning preposition in the PP as in *fclass_neg_subj_eff_reflobja_prepobj[um]_verb_neg* (in this case "um" (for); note the encoded reflexive direct object), we leave the pure syntax level.

As mentioned above, one of the goals is the combination of the resources (polarity lexicon and verb annotation). This combination provides us with new target specific sentiment calculations which were not possible in a compositional sentiment analysis purely relying on lexical resources and cannot be reliably inferred via a fuzzy criterion like nearness to other polar words. The effects and expectations of an instantiated syntactic verb pattern in combination with bottom-up propagated and composed polarity can therefore be used to approach the goal of sentence-level sentiment analysis based on a deep linguistic analy-

sis. Furthermore our system offers a possibility to detect violations of expected polarities ("admire a deceitful friend"), i.e., if the bottom-up composed polarity and the effects or expectations coming from the verb frame have an opposite polarity (see (Klenner et al., 2014b) and (Hollenstein et al., 2014)).

As a side-effect of this combination of resources our system can be used in future on the one hand to improve the polarity lexicon through automatic detection of good candidates for the lexicon in the case of reoccuring words on polar expectation for grammatical functions (e.g. "threaten so. with X"; X has a negative polarity expectation, see (Klenner et al., 2014a) for a similar approach). On the other hand, new syntactic patterns in combination with specific verbs can also be detected for annotation in the case of reoccurring bottom-up composed polarity. This procedure as a whole can then be applied especially for gathering domain specific resources.

6 Pipeline Architecture

The documents of our text corpus are parsed, transformed to VISL format and then composition takes place. Targets are identified at that stage as well, and if they are assigned as an argument (e.g. subject) to a modelled verb frame, expectations or effects are asserted. A feature selector then operates on the VISL output, extracting and accumulating polar information (see the next section). Clearly, polar features seem to be better suited to predict the positive, negative or controversial polarity of a target than its neutral polarity. Since text classification has proved successful in document-level polarity classification (Pang et al., 2002), we defined a pipeline where the class probabilities of a text classifier form additional input features to a second classifier. Our hypothesis was that both approaches, text classification and classification on the basis of polar feature vector turn out to be complementary.

More technically, in the first step, a text classifier is trained and applied to our text corpus using 5-fold cross validation. The results of the (test) folds are merged and the class probabilities are extracted and kept as features for the next step - the target polarity classification based on feature vectors comprising prior polarities, phrase level polarities produced by sentiment composition etc. (see next section).

We have experimented with various machine learning algorithms and frameworks, including SVM, Naïve Bayes, Logistic Regression, k-nearest Neighbor. We compared the results of the Stanford classifier[4] to those of Mallet, Megam and Rainbow. We found that Rainbow (McCallum, 1996) produced the best results for our text classification needs. On the other hand, Simple Logistic Regression as provided by Weka (Hall et al., 2009) performed best given our combined feature set. We experimented with feature selection, but none of the feature lists produced were able to outperform the class-specific feature selection automatically carried out by Simple Logistic Regression (cf. (Sumner et al., 2005)).

7 Feature Extraction

We have developed a feature extraction pipeline that extracts information about various polarity levels in words, phrases and sentences of the newspaper articles in our data set. Our feature selection chooses five sets of features which are then combined with the probabilities of the Rainbow text classification system to train a Simple Logistic classifier. With this method we allow features based on ordinal text classification as well as features based on our sentiment analysis resource.

In order to use our sentiment composition approach for machine learning we extract five different sets of features, resulting in a total of 150 features.

In short, our features are constructed as follows (referred to in Table 3):

1. *Text classification probabilities (Rainbow) (8 features)*: We take the output probabilities of Rainbow for each text as features for training the Simple Logistic classifier.

2. *Lexicon-based features (26 features)*: On the one hand, these comprise simple frequency counts of positive and negative words in the documents, taking into account the fine-grained information provided in our polarity lexicon. This means that we extracted

[4] *http://nlp.stanford.edu/software/classifier.shtml*

additional special features which are only concerned with the factual, moral or emotional values of the polar words in the training documents (as described in section 4), e.g. the sum of morally negative adjectives and nouns. On the other hand, we also include features capturing positive-negative ratios mapped to various dimensions. Moreover, we represent structural information by extracting features oriented at the title and the lead of the newspaper articles.

3. *Composition-based features (15 features)*: This feature set describes the information found in nominal and prepositional phrases mapped to the functional heads. Once more, it is possible to distinguish between features which represent frequency counts and features which represent polarity ratios.

4. *Verb-specific features (20 features)*: The goal of the verb-specific features is to extract the information modelled by our verb resource. For instance, we sum all occurrences of subjects and direct objects that receive a positive/negative "effect" from a verb. These features include the "effects" and "expectations" of a given verb as well as the polarity of the verb itself. Furthermore, we model the ratio between polar verbs and the amount of tokes in a text as well as the ratio between positive and negative verbs. These ratios can also be found in the lexicon-based and composition-based feature sets.

5. *Target-specific features (81 features)*: This last feature set is the largest one as it contains all of the information presented in the previous feature sets (2.)-(4.) in connection with phrases or sentences that include a target mention, e.g. the frequency of sentences in which a polar verb that has a direct relation to the target, or the frequency of a target appearing in a polar nominal or prepositional phrase. We also included different positive-negative ratios such as the ratio between targets which appear inside a positive phrase and targets which appear inside a negative phrase. Finally, we combined all the target-related features into two features which represent the complete amount of positive/negative information in the target sentences of one document.

We trained a Simple Logistic classifier on the described set of 150 features. Remarkably, fewer features reduced performance, although Simple Logistic always selected a proper subset of the features.

The impact of the five feature sets and the improvements achieved in comparison to the baseline system will be discussed in the next section.

8 Experiments

In our experiments, we seek to clarify three questions. What is the effect of polar features on classification accuracy? Does this effect depend on the text domain (e.g. finance versus insurance) and can we build high-precision classifiers by filtering text classification results accordingly?[5]

Articles	neut	neg	pos	contr	Entropy
5,000	0.18	0.36	0.19	0.28	0.584
10,000	0.35	0.28	0.14	0.22	0.580

Table 2: Class distribution.

8.1 Experiment I

In order to find out how strong the contribution of our new polarity resources and the features derived from it are, we draw a 5,000 document subset from the text corpus that maximizes target-verb-linkages. If a target is assigned as an argument to one of our verbs (e.g. is the subject or object of the verb), it inherits often a polarity (an effect or an expectation). Thus, the more such dependency links are found in a document, the stronger the impact should be. In other words, is it reasonable to extend our verb resource? Does it help to improve accuracy? Or is the performance independent of the applicability (the fitness) of our resource? We compared the results for the 5,000 set to a second subset comprising 10,000 documents, randomly drawn, but adhering to the distribution of the whole population (see Tab. 2).

[5] Our results in section 8.1 as well as the domain-specific results in section 8.2 based on accuracy all proved significant under the McNemar's paired test.

Description	5,000 articles					10,000 articles				
Feature sets	**Acc**	**neut**	**neg**	**pos**	**contr**	**Acc**	**neut**	**neg**	**pos**	**contr**
Baseline	50.02	34.2	62.8	49.6	**39.4**	52.32	62.2	56.7	37.1	**33.4**
Rainbow	49.86	34	63.1	49.7	37.7	52.58	62.2	57.4	37.2	32.1
Lexicon-based	50.98	33.6	64.5	53.4	36.7	52.66	62.7	57.3	37.7	30.6
Composition-based	50.78	33.1	64.6	53.0	36.5	52.75	62.9	57.3	37.7	30.6
Verb-specific	51.30	32.4	65.4	53.9	36.9	52.89	62.9	57.6	37.2	31.2
Target-specifc	**51.78**	**35.3**	**65.6**	**55.0**	36.7	**53.32**	**63.2**	**58.2**	**38.8**	31.8

Table 3: Results for dataset with 5,000 and 10,000 articles showing overall accuracy and f-measures for each class.

Description		Accuracy			Class Distribution			
Domain	**Articles**	**SA150**	**TC**	**TC+SA150**	**pos**	**neg**	**neut**	**contr**
Retail trade	1515	41.45	42.13	44.82	0.27	0.22	0.29	0.23
Pharma	3845	41.45	48.64	49.67	0.24	0.28	0.24	0.24
Transport	3155	44.98	48.54	50.11	0.13	0.33	0.27	0.27
Media	1310	46.64	47.25	50.53	0.11	0.35	0.27	0.27
Telecom	1438	48.09	51.02	50.54	0.19	0.24	0.38	0.19
Industry	1476	45.94	52.31	54.13	0.33	0.20	0.27	0.21
Insurance	4983	47.56	54.56	56.74	0.19	0.25	0.37	0.19
Banks	31373	51.94	61.43	63.34	0.12	0.43	0.36	0.18
Political inst.	3110	60.03	65.03	65.03	0.07	0.16	0.59	0.17
Unions	3685	71.46	72.20	73.33	0.05	0.11	0.67	0.17

Table 4: Domain-specific sentiment analysis (TC = Text Classification, SA150 = 150 sentiment analysis features).

In Tab. 3, the baseline (label Baseline) is taken from the output of rainbow (its class decision). We took also the class probabilities of rainbow as features (label Rainbow), followed by our polar features as described in the previous section. The improvement in accuracy is moderate (from 50.02% to 51.78%). However, those classes that should profit most from our features, namely negative and positive, actually do show a clear improvement: from 62.8% to 65.6% (negative) and from 49.6% to 55% (positive).

The baseline in accuracy on the right-hand side of Tab. 3 (10,000 texts) is higher (52.32% compared to 50.02%). However, the impact of our features is lower (1% giving 53.32%). Especially the impact on positive and negative classes is lower compared to the 5,000 subset which maximizes fitness of (our) resources.

Note that in both scenarios the (text classification) baseline accuracy of "controversial" decreases as our features are added. As mentioned in the introduction, we cannot deal with this kind of target evaluations, currently.

8.2 Experiment II

We wanted to know whether the classifier performance is stable in different domains, i.e. whether our resources and system components establish a (more or less) domain-independent machinery. We grouped the texts into their domains (e.g. finance, insurance etc.) and run the classifier. Tab. 4 shows that while the text classifier (TC) sets a different baseline depending on the domain (e.g. 42.13% Retail Trade; 72.20% Unions), the contribution of the polar features (TC+SA150) remains, compared to baseline variance, constant: the mean improvement is 1.4% (incl. one accuracy drop and one constant value). Note that in this experiment the full dataset is used (80,000). This explains performance drop compared to the (deliberately chosen) well fitting 5,000 subset.

There is one domain where performance stays constant (Political institutions) and one where it drops (Telecom). In both cases the majority class

is neutral, indicating, again, that our polar features better capture positive and negative than neutral and controversial cases.

Tab. 4 also shows the performance of the two classifiers independently from each other (SA150 compared to TC). We can see that text classification always produces higher values. For instance, for Retail trade: 41.45% compared to 42.13%. Since the sum of neutral and controversial (except for one case) together forms the majority of documents (see Tab. 4: Class distribution), this might just be a reflection of the slightly biased data (SA150 is good with positive and negative classes).

Since we have included a text classifier in our pipeline whose accuracy correlates with the probability of the decision (i.e. the confidence value), we wanted to know if we could create a scenario where we only give a classification for cases, where a certain probability is reached – implicating that accuracy would then also increase or at least not decrease. This scenario faces the challenge of the *fög* to cope with large amounts of newspaper articles every day. It is not only expensive to have human annotators classify the data, it might also be ineffective, since choosing a random sample of texts is always in danger of flaws concerning the representativeness of the sample. A high precision system would allow the *fög* to search for interesting texts, either from one of the classes, or even w.r.t. the polar load of texts.

As a further precondition, we set the minimum of the percentage of the documents that have to be classified (this number naturally decreases if one uses the probability of the classifier as a threshold) to 80%. Then we determine the concerning confidence value threshold and tried the classifier without and with our sentiment features only for those documents. It has to be noticed, that the high percentage of processed articles could only be reached with a Naïve Bayes (NB) classifier since the Maximum Entropy classifier (rainbow) had only high probabilities (relative to all probabilities in connection with good accuracy) for very small percentages.

Tab. 5 shows that this time the boost in accuracy when adding the sentiment features for the classification task is relatively stable over several domains. We can see that there is a gain in using the sentiment features along with the text classifier for the task even if "most difficult" cases for the text classifier are filtered out. This means that the improvement through the sentiment features does not only occur in the cases where the text classifier itself has decided badly.

Domain	**NB**	**NB+SA150**	**% articles**
Banks	60.97%	62.03%	90.4%
Pol. inst.	63.49%	64.23%	96.9%
Unions	72.8%	73.1%	96.5%
Insurance	55.7%	57.4%	90.6%
Transport	48.82%	50.96%	84.7%
Pharma	49.18%	49.96%	88.2%

Table 5: Results for different domains, filtered by probability of the text classifier (NB = Naive Bayes text classification, SA150 = 150 sentiment analysis features, % articles = percentage of articles processed under the corresponding accuracy).

9 Conclusion

We have introduced an approach for target-specific sentiment analysis that combines the output of a text classifier with features derived from fine-grained, compositional sentiment analysis. These two components are (at least in part) complementary: text classification better deals with class-specific wording (e.g. words indicating contrastive language), while polarity-based features better capture (and aggregate) the polar load of target-specific descriptions.

Our experiments have shown that operationalization of a class like "controversial" is difficult since there is no clear borderline to news texts which are slightly polar (positive, negative) or neutral. This is reflected in the fact that even human annotators reach only a poor interannotator agreement. Maybe a level of polarity (positive, negative) in combination with a single measurement for controversy could provide more reliable results since the (somehow subjective) decision could then be left to human judgement or to ex-post definitions.

The experiment with articles concerning different domains have shown some remarkable differences in the results. The baseline set by the text classifier varies considerably, whereas the contribution of our polar features is more or less stable.

This seems to indicate that the performance of text classification is much more domain-specific than features based on sentiment composition (and a general polarity lexicon).

Our experiments with a data subset of 5,000 texts that maximizes fitness of our resources have shown that the contribution of our features actually improve results on the proper polar classes, namely positive and negative. This is good news, since performance gain can now be coupled to the further development of our resources, especially the verb resource. However, especially with respect to the controversial dimension an in-depth error and data analysis is needed. We also hope to improve our evaluation process by creating more fine-grained annotated text, i.e., with annotation of certain text areas which lead the human annotator to his judgement relating to a specific target.

Acknowledgments

We would like to thank the *fög* institute, especially Mario Schranz and Urs Christen, for allowing us to use their huge target-level annotated text corpus. Also, we would like to thank the reviewers for their helpful comments and suggestions.

References

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity word sense disambiguation. In *EMNLP*, pages 190–199.

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2010. Sentiment analysis in the news. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 2216–2220, Valletta, Malta, May.

Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng. 2010. Identifying controversial issues and their sub-topics in news articles. In Hsinchun Chen, Michael Chau, Shu-hsing Li, Shalini Urs, Srinath Srinivasa, and G.Alan Wang, editors, *Intelligence and Security Informatics*, volume 6122 of *Lecture Notes in Computer Science*, pages 140–153. Springer Berlin Heidelberg.

Simon Clematide, Stefan Gindl, Manfred Klenner, Stefanos Petrakis, Robert Remus, Josef Ruppenhofer, Ulli Waltinger, and Michael Wiegand. 2012. MLSA - a multi-layered reference corpus for German sentiment analysis. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3551–3556.

David L. Deephouse. 2000. Media reputation as a strategic resource: An integration of mass communication and resource-based theories. *Journal of Management*, 26(6):1091–1112.

Shiri Dori-Hacohen and James Allan. 2013. Detecting controversy on the Web. In *CIKM '13*.

Mark Eisenegger and Kurt Imhof. 2008. The true, the good and the beautiful: Reputation management in the media society. In Betteke van Ruler Ansagr Zerfass and Sriramesh Krishnamurthy, editors, *Public Relations Research: European and International Perspectives and Innovation*. VS Verlag für Sozialwissenschaften.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Hollenstein, Nora and Amsler, Michael and Bachmann, Martina and Klenner, Manfred. 2014. SA-UZH: Verb-based Sentiment Analysis. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland.

Manfred Klenner, Michael Amsler, and Nora Hollenstein. 2014a. Inducing Domain-specific Noun Polarity Guided by Domain-independent Polarity Preferences of Adjectives. In Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2014), Baltimore, USA.

Manfred Klenner, Susanna Tron, Michael Amsler, and Nora Hollenstein. 2014b. The Detection and Analysis of Bi-polar Phrases and Polarity Conflicts. In Proceedings of the 11th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 14), Venice, Italy.

Isa Maks and Piek Vossen. 2012. A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53(4):680–688.

J. R. Martin and P. R. R. White. 2005. *Appraisal in English*. Palgrave, London.

Andrew Kachites McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow.

Karo Moilanen and Stephen Pulman. 2007. Sentiment composition. In *Proc. of RANLP-2007*, pages 378–382, Borovets, Bulgaria, September 27-29.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. Semantically distinct verb classes involved in sentiment analysis. In Hans Weghorn and Pedro T. Isaías, editors, *IADIS AC (1)*, pages 27–35. IADIS Press.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. *CoRR*, cs.CL/0205070.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proc. of HLT-EMNLP-05*, pages 339–346, Vancouver, CA.

Kevin Reschke and Pranav Anand. 2011. Extracting contextual evaluativity. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 370–374.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A new hybrid dependency parser for German. In *Proc. of the German Society for Computational Linguistics and Language Technology*, pages 115–124.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pages 1631–1642, Seattle, USA.

Marc Sumner, Eibe Frank, and Mark A. Hall. 2005. Speeding up logistic model tree induction. In Alípio Jorge, Luís Torgo, Pavel Brazdil, Rui Camacho, and João Gama, editors, *PKDD*, volume 3721 of *Lecture Notes in Computer Science*, pages 675–683. Springer.

Mikalai Tsytsarau, Themis Palpanas, and Kerstin Denecke. 2010. Scalable discovery of contradictions on the Web. In Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti, editors, *WWW*, pages 1195–1196. ACM.

VISL-group. 2013. *VISL CG-3. http://beta.visl.sdu.dk/cg3.html*. Institute of Language and Communication (ISK), University of Southern Denmark.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. of HLT/EMNLP 2005*, Vancouver, CA.